Multisensory integration in speech processing

Kilian-Hütten, N.; Formisano, Elia; Vroomen, J.

Published in:
Neural mechanisms of language

DOI:
10.1007/978-1-4939-7325-5_6

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

Chapter 6
Multisensory Integration in Speech Processing: Neural Mechanisms of Cross-Modal Aftereffects

Niclas Kilian-Hütten, Elia Formisano, and Jean Vroomen

6.1 Introduction

6.1.1 Multisensory Integration

Traditionally, perceptual neuroscience has focused on unimodal information processing. This is true also for investigations of speech processing, where the auditory modality was the natural focus of interest. Given the complexity of neuronal processing, this was a logical step, considering that the field was still in its infancy. However, it is clear that this restriction does not do justice to the way we perceive the world around us in everyday interactions. Very rarely is sensory information confined to one modality. Instead, we are constantly confronted with a stream of input to several or all senses and already in infancy, we match facial movements with their corresponding sounds (Campbell et al. 2001; Kuhl and Meltzoff 1982). Moreover, the information that is processed by our individual senses does not stay separated. Rather, the different channels interact and influence each other, affecting perceptual interpretations and constructions (Calvert 2001). Consequently, in the last 15–20 years, the perspective in cognitive science and perceptual neuroscience

N. Kilian-Hütten
Department of Psychiatry, Columbia University College of Physicians and Surgeons, New York, NY, USA

Department of Cognitive Neuroscience, Maastricht University, Maastricht, The Netherlands
e-mail: n.kilianhuetten@gmail.com

E. Formisano
Maastricht Brain Imaging Center, Maastricht University, Maastricht, The Netherlands
e-mail: e.formisano@maastrichtuniversity.nl

J. Vroomen (✉)
Department of Cognitive Neuropsychology, Tilburg University, Tilburg 5000 LE, The Netherlands
e-mail: J.Vroomen@uvt.nl

M. Mody (ed.), *Neural Mechanisms of Language*, Innovations in Cognitive Neuroscience, DOI 10.1007/978-1-4939-7325-5_6

has shifted to include investigations of such multimodal integrative phenomena. Facilitating cross-modal effects have consistently been demonstrated behaviorally (Shimojo and Shams 2001). When multisensory input is congruent (e.g., semantically and/or temporally) it typically lowers detection thresholds (Frassinetti et al. 2002), shortens reaction times (Forster et al. 2002; Schröger and Widmann 1998), and decreases saccadic eye movement latencies (Hughes et al. 1994) as compared to unimodal exposure. When incongruent input is (artificially) added in a second modality, this usually has opposite consequences (Sekuler et al. 1997).

6.1.2 Audiovisual Speech Perception

It becomes increasingly clear then, that in the case of spoken communication, the auditory modality is not of exclusive relevance. Indeed, visual speech signals in the form of lip movements, head movements, and gestures exert a significant influence on the perception of the auditory signal (Ross et al. 2007; Sumby and Pollack 1954; von Kriegstein 2012). It has been shown that extra-oral movements of the speaker's face and head correlate with the fundamental frequency and the amplitude of the speaker's voice (Munhall and Buchan 2004; Yehia et al. 2002) and that the addition of this information improves intelligibility (Munhall et al. 2004). However, the most informative visual aspect of speech is movement of the articulators. Over 60 years ago, it was demonstrated that processing of auditory speech signals improves when lip movements are visible (Sumby and Pollack 1954). This is especially relevant when the auditory signal is degraded or there is a substantial amount of overlapping, irrelevant signal. This can be easily appreciated in the noisy environment of a cocktail party or a poster session at a scientific conference, where one often focuses one's eyes more on the interlocutor's mouth to aid comprehension. Lip-reading is beneficial for auditory detection (Reisberg et al. 1987) and comprehension (Macleod and Summerfield 1990; Middelweerd and Plomp 1987; Sumby and Pollack 1954) and leads to improved performance equivalent to an increase in auditory signal-to-noise between 4 dB and 6 dB (Macleod and Summerfield 1990; Middelweerd and Plomp 1987) or even 12–15 dB (Sumby and Pollack 1954). It appears that these effects are present even in the absence of auditory noise (Remez 2012) and are strongest in moderate levels of noise (Ma et al. 2009; Ross et al. 2007).

Enhancement effects as discussed above reflect the most common situations: Here, auditory and visual channels are congruent and provide largely redundant information. In artificial situations, though, information sources can be put in conflict with each other: In the case of conflicting information, seen speech can actually alter both perceived location and identity of an auditory signal. In the ventriloquism effect, auditory and visual stimuli are presented synchronously, but shifted in space. This leads to a perceived displacement of the auditory source toward the visual one (Bertelson and Radeau 1981; Radeau and Bertelson 1977). The relative dominance of the individual modalities in this effect depends on the reliability (the inverse estimate of the noisiness) of each information source (Alais and Burr 2004), i.e.,

when the visual source is blurred, and thus poorly localized, the auditory modality dominates and influences the perceived location of the visual source. It has further been demonstrated that the ventriloquism illusion is not only effective in the spatial, but also in the temporal domain, where the perceived timing of a visual stimulus (e.g., a flash) can be biased toward an asynchronous sound (Vroomen and de Gelder 2004).

Besides perceived location or timing, conflicting inter-modal stimulation can even alter the perceived identity of auditory speech sounds. In a seminal paper that has now been cited more than 3100 times, McGurk and MacDonald (1976) demonstrated a powerful effect, in which lip-reading alters the percept of an auditory phoneme. When presenting a visual /ga/ together with an auditory /ba/, the resulting percept is /da/ (the McGurk illusion). It seems like the conflicting input results in a best-guess perceptual interpretation. The effect is extremely robust and has been a subject of intensive investigation since its discovery.

As a result of the behavioral findings discussed above, the last couple of decades has seen the emergence of a wealth of research tapping into the neural mechanisms of cross-modal integration, audiovisual speech perception and enhancement, as well as perceived-location and identity effects (see Calvert et al. 2004; Murray and Wallace 2011).

6.2 Audiovisual Speech Perception: Neural Mechanisms and Theories

6.2.1 The Modular View of Audiovisual Integration

Besides sub-cortical structures—predominantly the superior colliculus (SC)—and regions traditionally regarded as unisensory, such as early auditory and visual regions, visual area MT (middle temporal; responsible for the processing of motion), or the Fusiform Face Area (FFA), neuroimaging studies have identified a network of higher-order integrative areas that are involved in the processing of audiovisual speech. More specifically, middle and posterior superior temporal sulcus (pSTS), Broca's area, dorsolateral prefrontal cortex, superior precentral sulcus, supramarginal gyrus, angular gyrus, intraparietal sulcus (IPS), inferior frontal gyrus (IFG), and insula have all been implicated in these processes (Beauchamp et al. 2004a, b; Callan et al. 2001, 2003, 2004; Calvert et al. 1999, 2000, 2004; Capek et al. 2004; Miller and D'esposito 2005; Möttönen et al. 2004; Olson et al. 2002; Pekkola et al. 2005). At least for the homologues of pSTS and IPS, neurons have been found in non-human primates that respond to both auditory and visual stimulation (Ghazanfar and Schroeder 2006). Furthermore, some of these regions have been shown to be activated by silent speech reading. This is true for middle and posterior superior temporal sulcus (Callan et al. 2004; Ludman et al. 2000; MacSweeney et al. 2002; Skipper et al. 2005), inferior frontal gyrus and Broca's area (Campbell et al. 2001;

Ojanen et al. 2005; Watkins et al. 2003), and possibly even for primary auditory cortex (Calvert et al. 1997; Pekkola et al. 2005). Most studies reveal a left-over-right hemisphere asymmetry in activation (Capek et al. 2004).

Classical studies in cats' superior colliculi (Stein et al. 1988; Stein and Meredith 1990) first identified multisensory neurons that exhibit supra-additive firing patterns in response to multisensory, as compared to unisensory, input (AV > A + V). Early functional magnetic resonance imaging (fMRI) studies consequently searched for brain regions whose hemodynamic response (blood-oxygen level dependent response; BOLD) mimicked this activation pattern (Calvert et al. 2000). Since several studies failed to replicate successful results using this, quite strict, criterion (Beauchamp et al. 2004a, b; Beauchamp 2005; Laurienti et al. 2005; Stevenson et al. 2007), other criteria, such as a multimodal response that is stronger than the stronger one of the two unimodal responses, or inverse effectiveness (multisensory enhancement should *increase* as a function of stimulus quality degradation) have been applied. Most of these criteria have received a substantial amount of criticism. For a discussion of this, see Laurienti et al. (2005), James and Stevenson (2012), and Stein et al. (2009).

In spite of the criticism, left pSTS has been repeatedly and robustly implicated in audiovisual integration on the basis of these criteria (Beauchamp et al. 2004a, b; Calvert et al. 2000; Miller and D'esposito 2005; Nath et al. 2011; Stevenson and James 2009; Wright et al. 2003). Furthermore, when audiovisual stimuli are incongruent, usually a depression of activity in this region results (Campbell and Capek 2008; Wright et al. 2003) and, as mentioned before, left pSTS is activated by both audiovisual speech and by silent speech-reading (Callan et al. 2004; Campbell and Capek 2008; Capek et al. 2004; Hall et al. 2005; MacSweeney et al. 2002; Skipper et al. 2005) and differences in left STS activation have been related to language comprehension (McGettigan et al. 2012). More recently, results from electrocorticography suggest a dissociation between anterior and posterior STG in responses to audiovisual speech with clear vs noisy auditory component. The pSTG not aSTG appears to be important for multisensory integration of noisy auditory and visual speech (Ozker et al. 2017). Also, single-pulse transcranial magnetic stimulation (TMS) over pSTS has been shown to disrupt the perception of McGurk effects in a time window from 100 ms before the onset of the auditory stimulus to 100 ms after onset (Beauchamp et al. 2010).

6.2.2 Multisensory Processing as the Default Mode of Speech Perception

The findings discussed so far follow the traditional modular view that assumes that multisensory integration takes place only in higher-order multisensory regions after extensive unisensory processing in the respective cortices. However, more recently, this view has been challenged (Ghazanfar and Schroeder 2006; Schroeder et al. 2008). Increasing evidence is accumulating for the idea that multisensory processing can be regarded as the default mode of speech perception and that the

integration of auditory and visual speech signals already occurs in the earliest stage of processing in, presumptively unisensory cortical areas (Driver and Noesselt 2008; Ghazanfar and Schroeder 2006; Ghazanfar 2012; Rosenblum et al. 2005). Studies in non-human primates (Ghazanfar et al. 2005, 2008; Kayser et al. 2005, 2007; Lakatos et al. 2007), as well as in humans (Besle et al. 2004, 2009; Pekkola et al. 2005; Stekelenburg and Vroomen 2007; Van Wassenhove et al. 2005; Vroomen and Stekelenburg 2010) have demonstrated the integrative influence of visual (including speech signals) and somatosensory signals on auditory processing in primary and lateral belt auditory cortex. It has been suggested that the underlying mechanism of such cross-modal modulation of early auditory cortical processing may be based on a predictive resetting of the phase of the ongoing oscillatory cycles of neuronal ensembles (Schroeder et al. 2008).

Several accounts for these early cross-modal influences on presumptively unisensory cortices exist, which offer different, although not necessarily mutually exclusive, explanations for the origin of these modulations (Driver and Noesselt 2008; Schroeder et al. 2003). More specifically, it is conceivable that (a) all cortical regions are essentially multisensory and receive input from thalamocortical pathways and direct cortico-cortical connections between different sensory cortices (Ghazanfar and Schroeder 2006); (b) that there is still a separation between unisensory and multisensory integration areas, but new convergence zones exist earlier in the hierarchy and closer to unisensory regions than previously assumed (Beauchamp et al. 2004a, b); and (c) that cross-modal modulations of sensory-specific cortical processing reflects feedback influences from multisensory convergence zones (Driver and Noesselt 2008; Jiang et al. 2001). While account (b) basically amounts to a new parcellation of cortex within the traditional perspective, account (a) represents a rather extreme new view on cortical processing. Support for account (a) comes from neuroanatomical studies demonstrating the involvement of thalamocortical (Cappe et al. 2009; Hackett et al. 2007; Lakatos et al. 2007) and monosynaptic cortico-cortical connections between primary auditory and primary visual cortex (Clavagnier et al. 2004; Falchier et al. 2002, 2010), from direct connections between voice- and face-processing areas (Blank et al. 2011), and from reports of very early post-stimulus cross-modal influences on the event-related potential (ERP; within approximately 50 ms) (Giard and Peronnet 1999; Molholm et al. 2002; Senkowski et al. 2007). However, feedback connections from convergence zones, such as pSTS, still seem to clearly outnumber direct connections between early sensory-specific cortices (Falchier et al. 2002). Also, it is unclear whether the information transmitted along this route is stimulus- or percept-specific and, thus, reflects actual multisensory integration, or whether it represents more general modulations, such as attention or arousal effects (Driver and Noesselt 2008). Furthermore, ERP studies demonstrating extremely early effects based on the additive model have been criticized for not taking into account common, non-specific activity, such as attention effects and anticipatory, and motor, responses (Cappe et al. 2010; Gondan and Röder 2006; Teder-Sälejärvi et al. 2002). Controlling for these factors typically delays the effects to approximately 60–100 ms. Over the last decade it has been shown that non-linear multisensory interactions in the ERP follow from topographic

modulations, and result in sub-additive responses that are functionally coupled within primary auditory and visual cortices, as well as pSTS (Cappe et al. 2010). Support for account (c) the notion that multisensory effects in sensory-specific cortices reflect feedback influences from higher-order convergence zones comes from comparisons of latencies in the electroencephalogram (EEG) (Besle et al. 2004; Ponton et al. 2009), from fMRI studies investigating functional contrasts (Calvert et al. 2000), connectivity (Noesselt et al. 2007) and experimental differentiations between integration responses and perceptual effects (Kilian-Hütten et al. 2011a, b; Sohoglu et al. 2012), and from studies interfering with normal brain functioning in order to establish cause-and-effect relationships in cats (Jiang et al. 2001) and humans (Beauchamp et al. 2010).

It has become increasingly clear that the different accounts just discussed are not mutually exclusive, but that feedforward, lateral, and feedback connections exist and that multisensory integration involves all of these, relying more or less on particular types of integration depending on stimulus and task context (Besle et al. 2008, 2009; Driver and Noesselt 2008; Schroeder et al. 2003).

6.2.3 Theoretical Accounts of Audiovisual Speech Perception

In the context of audiovisual speech perception, one important distinction here might be what has been termed correlation versus complementary mode (Campbell and Capek 2008). These two proposed modes of processing (correlation versus complementary) focus on the relation between auditory and visual speech stimuli and are based on the observation that auditory comprehension benefits from visual information in two ways; first, the auditory and the visual signal are highly correlated in terms of temporal dynamics and the speech processing system exploits these redundancies (correlation mode), and second, when the quality of the auditory signal is degraded, or certain speech segments are acoustically ambiguous, the visual signal can help disambiguate the acoustics and aid perception (complementary mode) (Campbell and Capek 2008). There is support for the idea that the specific locus of multisensory integration is affected by the relative importance of these modes in a particular experimental context. Callan et al. (2004) varied the visibility of facial detail using spatial filtering and could show that, while middle temporal gyrus (MTG) activation was increased when fine detail was accessible, pSTS activation was unaffected, lending evidence to the idea that pSTS is driven primarily by the dynamic aspects of the audiovisual speech stream (correlated mode), rather than by specific facial detail (complementary mode) (Campbell and Capek 2008).

The differences between correlation mode and complementary mode may reflect a more general phenomenon; reliability-based cue weighting (Fetsch et al. 2012; Nath and Beauchamp 2011; Sheppard et al. 2013). It has been demonstrated repeatedly in behavioral studies that, in multisensory integration, subjects give stronger weighting to the more reliable modality and that these weightings are adapted in a dynamic, context-dependent fashion (Alais and Burr 2004; Ernst and Banks 2002;

Ma et al. 2009). This leads to an optimal solution, because it creates estimates with the lowest possible variance, which, in turn, results in superior perceptual performance as compared to what can be achieved based on either unisensory signal alone, or with any predetermined weighting pattern (Fetsch et al. 2012). In audiovisual speech perception, this may result in dynamic changes of functional connectivity between auditory and visual sensory cortices, respectively, and integration cortices, such as pSTS, depending on the reliability of the auditory and visual speech signals (Nath and Beauchamp 2011).

In the light of all these findings, it is vital from a computational perspective to understand how auditory recognition can benefit from visual input. One framework that may be applicable here is based on a theory that has been proposed to more generally account for perceptual inference (i.e., the recognition of perceived objects) and perceptual learning effects (Friston 2005). This scheme, referred to commonly as "predictive coding," rests on the idea that an integral part in perceptual inference is minimizing free energy, or more pragmatically, error. This is done by relying on a hierarchical model, where sensory responses at lower levels of the hierarchy are predicted at higher levels. In return, lower levels send prediction errors to higher levels, enabling learning. In other words, this idea relies on an empirical Bayesian model, where prior expectations can be formed, which in turn, exert their influence in the light of sensory evidence in a dynamic and context-dependent fashion. In the realm of multisensory integration, priors may originate in another modality (von Kriegstein 2012). For speech, it is conceivable that visual information (lip movement) biases processing in the auditory cortex, probably (but not necessarily) via feedback connections from higher-order convergence zones. Recently, support for this idea has been found in purely auditory speech (Clos et al. 2014), in auditory speech primed by written text (Sohoglu et al. 2012), and in audiovisual speech (Arnal et al. 2011; Noppeney et al. 2008). It has been suggested that, in an oscillation framework, higher frequencies (in the gamma (~30–60 Hz) and high-gamma (~70–80 Hz) frequency range) may be primarily involved in the signaling of bottom-up prediction errors, while slower frequencies (beta frequency range) would communicate top-down predictions (Arnal et al. 2011; Arnal and Giraud 2012).

6.3 Cross-Modal Aftereffects

6.3.1 Hysteresis Versus Adaptation

The discussion so far has focused on situations where the speech signals from both modalities (auditory and visual) are presented concurrently and affect each other directly. This is true for normal speech perception and also when one or both modalities are noisy, or when the perceived location and/or identity of a stimulus are altered on the fly. However, there are also cases where cross-modal effects can alter unisensory perception beyond the immediate presentation in time. Such aftereffects

have been found in unisensory phenomena, such as the waterfall illusion (Purkinje 1820) where after looking at a waterfall for an extended period of time, stationary rocks, for instance, seem to be moving upward, or the prism experiments by Stratton (1897) in which he found that a radical conflict between proprioception and visual field (which was turned upside down using special goggles) over time led to an adaptation of visual perception. Aftereffects can have two directions: Negative aftereffects like the waterfall illusion, the tilt aftereffect (the perceived change in orientation of a line or grating after prolonged exposure to another oriented line or grating; Gibson and Radner 1937), or color-opponency likely reflect a "fatigue" of neural sensors, while positive aftereffects, such as prism adaptation, likely reflect a "learning" of new sensory arrangements. Such attractive aftereffects (making a similar percept more likely) are also known as hysteresis, while repulsive aftereffects (making a similar percept less likely) are also referred to as adaptation.

In the multisensory domain, the ventriloquist illusion has been shown to produce attractive aftereffects after prolonged exposure (Bertelson et al. 2006; Radeau and Bertelson 1974, 1977). As was mentioned before, in this effect, the perceived location of a sound is shifted in space toward a synchronously presented visual target (Bermant and Welch 1976; Bertelson and Radeau 1981; Bertelson and Aschersleben 1998; Klemm 1909). The associated aftereffects are in line with this immediate effect—the perceived location of sounds presented in isolation after audiovisual exposure is shifted toward the visual stimuli presented during the exposure phase. In audiovisual speech perception, traditionally, contrastive effects have been found, evident in selective speech adaptation (Roberts and Summerfield 1981), where the repeated presentation of a nonambiguous phoneme leads to a decrease of the probability of reporting the same percept when tested with an ambiguous phoneme. In other words, repeated exposure to a nonambiguous /aba/ leads to a reduction of subsequent /aba/ perception, a phenomenon that, as mentioned before, may be explained by neuronal fatigue (Anstis et al. 1998; Eimas and Corbit 1973) (but see: Diehl et al. 1978; Diehl 1981; Samuel 1986).

6.3.2 The Initial Study on Audiovisual Recalibration of Auditory Speech Perception

Bertelson et al. (2003), however, were inspired by the findings of hysteresis effects in the ventriloquist illusion and investigated the possibility of similar aftereffects in the domain of audiovisual speech perception. They hypothesized that the crucial manipulation for achieving a hysteresis effect, as opposed to an adaptation effect, was the ambiguity of the auditory component of the adapter stimuli. While in the classical McGurk effect, as well as in selective speech adaptation paradigms, a nonambiguous phoneme is used, Bertelson et al. (2003) synthesized an ambiguous phoneme halfway between a nonambiguous /aba/ and a nonambiguous /ada/ (*A?*) and dubbed this sound onto videos of a speaker pronouncing /aba/ (A?Vb) or /ada/

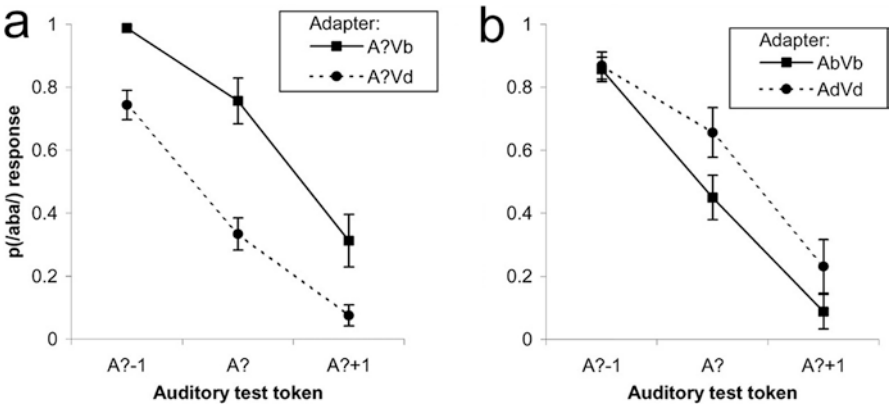

Fig. 6.1 The graphs show the proportion of /aba/ judgments after exposure to an adapter consisting of (**a**) the participant's ambiguous auditory token (A?) combined with either visual /aba/ (A?Vb) or visual /ada/ (A?Vd) or (**b**) a nonambiguous auditory token (Ab or Ad) combined with the congruent visual stimulus, /aba/ (AbVb) or /ada/ (AdVd). Figure (**a**) clearly shows a recalibration effect, while (**b**) indicates selective speech adaptation

(A?Vd), respectively. They showed that the repeated exposure to one type of video (A?Vb or A?Vd, respectively) increased the probability of corresponding perceptual interpretations in auditory-only post-tests. In the original study, eight videos were presented, followed by six post-tests (A? twice, plus the two tokens closest to it on the /aba/-/ada/ continuum; A? − 1 and A? + 1). This procedure was repeated a large number of times in random order. It could be shown that the proportion of /aba/ responses was significantly higher following A?Vb exposure than following A?Vd exposure (see Fig. 6.1a).

Crucially, when nonambiguous auditory tokens were dubbed onto the videos (AbVb/AdVd), the contrasting effect was found, i.e., selective speech adaptation (Fig. 6.1b). This is especially remarkable because subjects could not perceptually distinguish the ambiguous (A?Vb/A?Vd) from the nonambiguous (AbVb/AdVd) videos (Vroomen et al. 2004). This finding also rules out the possibility of an explicit strategy endorsed by the subjects, since they could not know, for a given block, whether they were exposed to ambiguous or nonambiguous stimuli.

6.3.3 Differences Between Recalibration and Selective Speech Adaptation

Besides the disparity in direction of effect, recalibration, and selective speech adaption also differ in a number of psychophysical characteristics; namely in buildup, dissipation, and the necessity of processing stimuli within a "speech mode" (Vroomen and Baart 2012).

In order to investigate the buildup courses of the two phenomena, Vroomen et al. (2007) presented continuous streams (up until 256 trials) of audiovisual exposure using the "ambiguous" recalibration adapters (A?Vb/A?Vd) and the "nonambiguous" adaptation adapters (AbVb/AdVd), respectively, and regularly inserted test trials. It was shown that selective speech adaptation effects linearly increased with the (log-) number of exposure trials, which fits with an accumulative fatigue idea. Recalibration, however, reached ceiling level already after eight exposure trials and then, surprisingly, fell off after 32 exposure trials with prolonged exposure. It was suggested that in this case, both recalibration and selective speech adaptation run in parallel and the latter dominates with prolonged exposure due to an increasing effect size.

Recalibration and selective speech adaptation also turned out to differ in terms of dissipation. Vroomen and de Gelder (2004) used a large number of exposure trials (50 trials of one kind), followed by 60 test trials. While the recalibration effect already dissipated after as little as six exposure trials, the effects of selective speech adaptation were still manifest even after 60 trials.

Lastly, Vroomen and Baart (2009) investigated the speech-specificity of both effects. In order to do so, they relied on sine-wave speech, a manipulation of speech stimuli which reduces the richness of the speech sound by removing all of its natural attributes and retaining only the pattern of vocal tract resonance changes; hence, these stimuli can be perceived either as speech or as non-speech, depending on the subject's perceptual mode. In order to manipulate perceptual mode, subjects were trained to distinguish two sine-wave stimuli as either /omso/ and /onso/ (speech mode) or as stimulus 1 and 2 (non-speech mode). It could be shown that recalibration crucially relies on being in speech mode (recalibration took place in speech mode, but not in non-speech mode), while perceptual mode had no effect on selective speech adaptation (which was effective in both modes).

6.4 The Neural Mechanisms of Cross-Modal Recalibration

The uniqueness of cross-modal recalibration makes it an intriguing subject for neuroscientific investigation. The reason for this is twofold; first, the aforementioned psychophysical differences suggested that cross-modal recalibration has a distinct neural substrate from that of selective speech adaptation (i.e., neuronal fatigue), rendering it crucial to find the origin of this perceptual biasing signal. Second, the recalibration paradigm yielded the unique possibility of comparing the neuronal responses underlying the differential perceptual interpretation (/ada/ or /aba/) of physically identical, auditory stimuli. In other words, recalibration made it possible to experimentally disentangle acoustics from perception and to examine the neuronal underpinnings of a purely perceptual difference.

In order to investigate these aspects of the recalibration phenomenon, Kilian-Hütten et al. (2011a, b) adapted the classical setup employed in the original Bertelson et al. (2003) study for the functional magnetic resonance imaging (fMRI) environment (Fig. 6.2). Subjects were presented with blocks of eight identical audiovisual

Fig. 6.2 Schematic overview of the paradigm used by Kilian-Hütten et al. (2011a, b), which was based on the behavioral study by Bertelson et al. (2003). Each run consisted of ten mini runs, which each comprised eight audiovisual exposure trials (A?Vd or A?Vb), followed by six auditory post-test trials. Audiovisual exposure was presented following a block design, while an event-related presentation scheme was applied for the auditory test trials

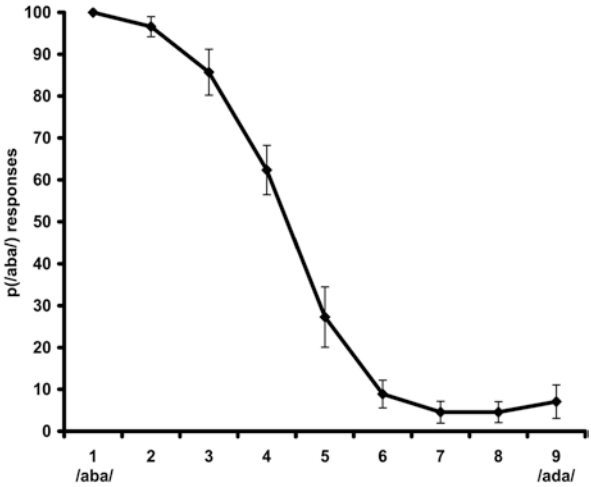

Fig. 6.3 Results of the auditory pretest. The mean proportion (p) of /aba/ classifications across the 11 participants for each stimulus on the nine-step /aba/-/ada/continuum are presented. Sound #4 was chosen as A? for eight of the participants and sound #5 for the remaining three

adapters (A?Vb and A?Vd, respectively), which were each followed by six auditory test trials consisting of forced choice perceptual judgments (/aba/ vs. /ada/) of ambiguous stimuli (A?, A? + 1, A? − 1). The most ambiguous stimulus on the nine-step continuum ranging from /aba/ to /ada/ was identified individually per subject in a pretest (Fig. 6.3). Behaviorally, the results of the original psychophysical study (Bertelson et al 2003) could be replicated (Fig. 6.4). Scanning was performed using a mixed block/event-related design, where audiovisual exposure trials were presented in blocks, while auditory test trials were presented in slow event-related fashion. This enabled the authors to achieve high signal-to-noise ratios for the exposure blocks, while preserving the possibility of analyzing auditory test trials on a trial-by-trial basis, depending on the subjects' perceptual judgments. This last point was crucial in order to allow for the possibility of investigating the neural substrate of the purely perceptually distinct categorization (/aba/ vs. /ada/) of physically identical stimuli (A?).

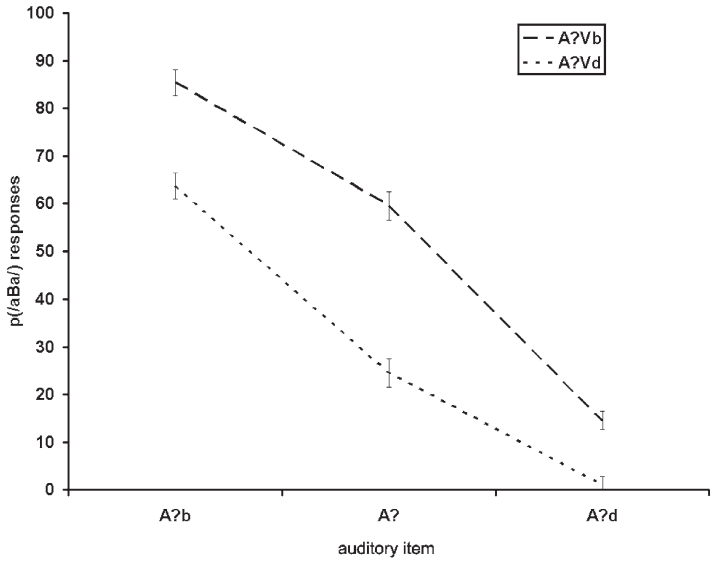

Fig. 6.4 Behavioral results of the auditory post-tests in Kilian-Hütten et al. (2011a, b). For the participant's ambiguous auditory item and its two neighbors on the continuum, the graph shows the proportion of /aba/ classifications after exposure to an audiovisual adapter comprised of the ambiguous item paired with either visual /aba/ (A?Vb) or with visual /ada/ (A?Vd). For all three auditory items, the difference in proportion /aba/ responses after exposure to the A?Vb adapter vs. the A?Vd adapter was significant

6.4.1 The Effects of Recalibration: Decoding Perceptual Interpretations of Ambiguous Phonemes

Traditionally, neuroimaging studies on perception have faced a confounding issue. This is because distinct percepts usually follow distinct physical stimuli. In other words, when /aba/ is presented, a subject perceives /aba/ and when /ada/ is presented, a subject perceives /ada/. Hence, when applying subtraction logic, as is traditional procedure in fMRI research, the two conditions that are being compared usually differ not only in percept, but also in physical stimulus characteristics. When one is interested in the neural substrate of perception proper, and not of stimulus processing per se, this is a problem. The recalibration phenomenon, together with event-related stimulus presentation, allowed disentangling auditory perception from stimulus acoustics.

Kilian-Hütten et al. (2011a) combined the recalibration paradigm with a machine learning approach in order to decode auditory perception on a trial-by-trial basis from the fMRI signal. This approach, also referred to as multivoxel pattern analysis (MVPA), applies machine learning methods to the multivariate analysis of fMRI data sets (Formisano et al. 2008; Haxby et al. 2001; Haynes and Rees 2005a, b). A pattern classification algorithm is typically trained, using a large set of trials, to

associate a given experimental condition or cognitive state with a distributed pattern of fMRI responses. The trained classifier can then be tested on new, unseen fMRI patterns to decode the associated cognitive state. In other words, rather than predicting brain responses from experimental conditions, as in the general linear model (GLM; a regression analysis using regressors based on the timing of experimental conditions), MVPA can "predict" the experimental condition from the brain response, which is why it has often been denoted as a "brain reading" approach. One particular class of MVPA algorithms are support vector machines (SVM).

Kilian-Hütten et al. (2011a) trained an SVM to learn the association between multivariate patterns of fMRI signal and corresponding labels, determined by the subjects' perceptual interpretations. In other words, while the physical stimuli were identical for both labels, the percept differed, and the SVM was trained to decode the respective percept on a trial-by-trial basis from the fMRI signal. The analysis was anatomically confined to the temporal lobes in order to test the hypothesis that abstract auditory representations can be found in early auditory cortex. Besides accuracy levels (which were significantly above chance, as validated with permutation testing), it is interesting to visualize the spatial activation patterns that were used for classification. To this end, group discriminative maps, i.e., maps of the cortical locations that contributed most to the discrimination of conditions, were created after cortex-based alignment (Goebel et al. 2006) of single-subject discriminative maps (Staeren et al. 2009). For SVM analyses it is meaningful, at an individual map level, to rank the features (i.e., voxels) relatively according to their contribution to the discrimination. In the resulting group-level discriminative maps, a cortical location (vertex) was color-coded if it was present among the 30% of most discriminative vertices in the corresponding individual discriminative maps of at least seven of the 11 subjects. As can be seen in Fig. 6.5, these maps identified left-lateralized clusters of vertices along the posterior bank of Heschl's gyrus, Heschl's sulcus, and, adjacently, in the anterior portion of planum temporale (PT). Additional clusters of smaller extent were found at the left temporoparietal junction and, bilaterally, on middle superior temporal gyrus (STG) and sulcus (STS).

These results showed that pure perceptual interpretation of physically identical phonemes can be decoded from activation patterns in early auditory cortex. Thus, beyond the basic acoustic analysis of sounds, constructive perceptual information is present in regions within the anterior PT, tangent to the posterior bank of Heschl's gyrus and sulcus.

6.4.2 The Origin of Recalibration: Predicting Recalibration Strength from Cortical Activation

The results just discussed above concentrated on the effects of cross-modal recalibration, i.e., the change in auditory perception. The obvious next question that arose was: Where does cross-modal recalibration itself take place and, thus, what is the

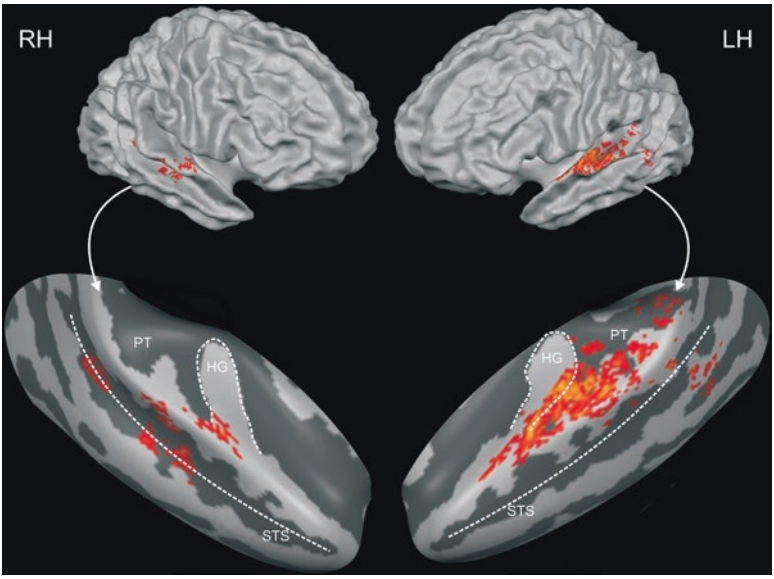

Fig. 6.5 Discriminative map from Kilian-Hütten et al. (2011a). Group map of the 30% of active voxels most discriminative for the purely perceptual difference between /aba/ and /ada/. A location was color-coded if it was present on the individual maps of at least seven of the 11 subjects. Maps are overlaid on the reconstructions of the average hemispheres of the 11 subjects (*top*) and on inflated reconstructions of the *right* and *left* temporal lobes of these average hemispheres (*bottom*). *RH* right hemisphere, *LH* left hemisphere, *HG* Heschl's gyrus

origin of the perceptually biasing effect? In order to answer these research questions, Kilian-Hütten et al. (2011b) focused their efforts on cortical activation during the audiovisual exposure trials. In a first step, a simple comparison between blocks of audiovisual exposure and baseline identified a network of brain regions corresponding to those generally found in audiovisual speech perception paradigms (primary and extrastriate visual areas, primary and early auditory areas, STG/STS, inferior frontal sulcus (IFS), premotor cortex, and inferior parietal lobe, touching upon angular and supermarginal gyri and the intraparietal sulcus). This was expected and replicated prior results. However, this does not mean that this whole network is responsible for the recalibration effect. In order to identify the subset of regions for which this is indeed true, the authors applied a behaviorally defined contrast. The strength of the recalibration effect is variable from one given exposure block to the next and can be quantified as the number of auditory post-tests perceived in line with the type of exposure block (A?Vb or A?Vd). These values could be employed to identify brain regions whose activation during the exposure blocks varied with the strength of the recalibration effect. The cortical activation in these areas, thus, predicted the perceptual tendency later in time. Beyond the basic identification of responsive regions, recalibration, thus, made it possible to pinpoint those regions which were responsive to audiovisual stimulation and which, further, were functionally relevant in driving the biasing perceptual effect. The network of regions for

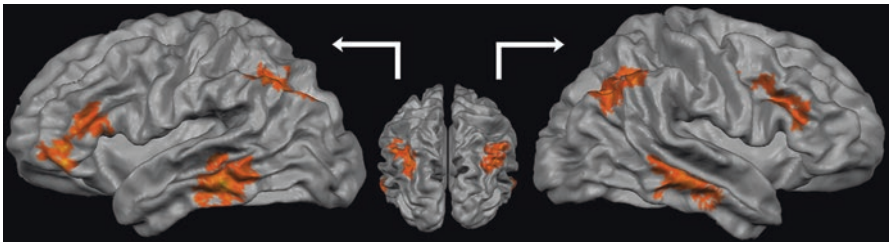

Fig. 6.6 Group results for the behaviorally weighted contrast used by Kilian-Hütten et al. (2011b) overlaid on the average hemispheres obtained from the cortex-based alignment procedure. Shown are bilateral IPL, IFS, and posterior middle temporal gyrus

which this was true included bilateral IPL, IFS, and posterior middle temporal gyrus (Fig. 6.6). These results were further corroborated by a functional connectivity analysis (psychophysiological interaction analysis; PPI), which demonstrated that this network of areas exhibits increased functional/effective connectivity with the left auditory cortex during blocks of audiovisual exposure relative to baseline.

These findings are in correspondence with the results from other investigations of perceptual learning (Myers et al. 2009; Naumer et al. 2009; Raizada and Poldrack 2007). Gilbert et al. (2001, p. 681) define perceptual learning as "improving one's ability, with practice, to discriminate differences in the attributes of simple stimuli." In the case of recalibration, the disambiguating information from audiovisual exposure biases auditory perception such that it can be regarded as improved in reference to the (momentary) demands of sensory reality. What seems to happen in the case of cross-modal recalibration, thus, is that integrative audiovisual learning effects take place in the identified network, which in turn affect later constructive (auditory) perceptual processes.

6.4.3 Theories of Audiovisual Speech Perception and the Neural Substrate of Recalibration

Taking together the results from the studies discussed above, a full neural model of cross-modal recalibration emerges (Fig. 6.7). A higher-order network including IPL, IFS, and MTG is suggested to process integrative learning effects, and consequently install a perceptual bias in auditory regions, most prominently the left Heschl's sulcus and the planum temporale, influencing future constructive auditory perception.

This interpretation of the results is in line with a model of the neural mechanisms of hysteresis and adaptation recently put forward by Schwiedrzik et al. (2014). In their exclusively visual study, they were able to dissociate hysteresis and adaptation effects in a single paradigm. What they found was that, while adaptation effects were largely confined to early visual areas, hysteresis effects mapped onto a more widespread and higher-order fronto-parietal network. Using a modeling approach, they showed that their results could be explained in a Bayesian framework.

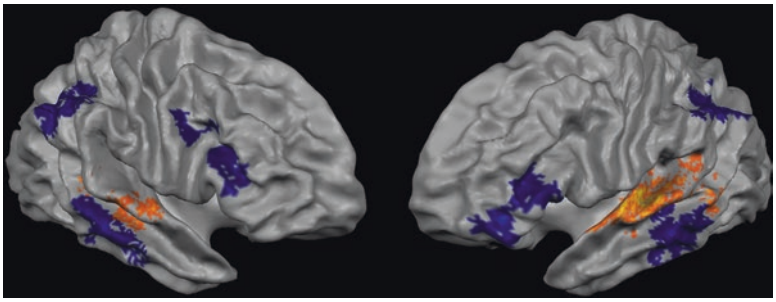

Fig. 6.7 The proposed model (Kilian-Hütten et al. 2011a, b). A higher-order network (in *blue*) including IPL, IFS, and MTG is suggested to process integrative learning effects (cross-modal recalibration), and consequently install a perceptual bias (the prior) in auditory regions (in *red* and *orange*), most prominently the left Heschl's sulcus and the planum temporale, influencing future perceptual interpretations of sensory input (the likelihood function), resulting in recalibrated auditory perception

A Bayesian approach is well-suited for this problem, because it takes into account both the available evidence (the likelihood function) and knowledge about the world (the prior). In terms of hysteresis and adaptation, this means that perception is determined by the sensory evidence (which is bimodal in the case of bistable stimuli) and by the prior. While the former is influenced by adaptation (neuronal fatigue), the latter is determined by hysteresis. More specifically, when a certain interpretation is primed, the prior is adjusted toward this interpretation, inducing hysteresis. This can be regarded as a special case of the general model of predictive coding. As discussed before, in this framework, sensory responses at lower levels of the hierarchy are predicted at higher levels. In return, lower levels send prediction errors to higher levels, enabling learning. Also Bayesian in nature, this model states that prior expectations can be formed and, in turn, exert their influence in the light of sensory evidence in a dynamic- and context-dependent fashion.

In the case of cross-modal recalibration, this would mean that the fronto-parietal network computes the prior and communicates the outcome to early auditory regions. Here, a representation of the perceptual interpretation can be decoded. This representation is almost exclusively determined by the prior, because the ambiguity of the stimulus creates a bimodal likelihood function, which has to be disambiguated by the prior before perception can take place.

6.5 Conclusion

Audiovisual speech perception involves the coordinated processing of a large network of brain regions. Besides unisensory cortices, higher-order regions are involved in the convergence and integration of multisensory input, as well as in the semantic and cognitive appraisal of this information. Cross-talk between the senses has an impact on perception, including perceived location and perceived identity of sensory input. Multisensory processing improves recognition accuracy particularly for speech

in noise through multiple stages on integration supported by distinct neuroanatomical mechanisms (Peelle and Sommers 2015). It seems that in order to optimally integrate auditory and visual information in the perception of speech, the brain exploits several connections, including feedforward, lateral, and feedback pathways. The relative importance of these connections depends on stimulus and task conditions. One important factor in this context is the relative reliability of the sensory input, as explained by reliability-based cue weighting and predictive coding frameworks.

These frameworks are also important in accounting for the neural bases of cross-modal perceptual aftereffects, such as recalibration. Recalibration is a hysteresis effect in that it elicits post-exposure perceptual biases that are in line with the percepts experienced during audiovisual exposure. This, along with differences in buildup, dissipation, and the necessity of processing stimuli in "speech mode," suggests that the neural substrate of this effect is different from that of selective speech adaptation (neuronal fatigue). Using a machine learning approach and behaviorally weighted GLM contrasts, it could be shown that cross-modal recalibration is reflected in integrative audiovisual learning effects that take place in a higher-order network involving IPL, IFS, and posterior middle temporal gyrus and which then install a perceptual bias in early auditory regions, impacting later auditory perception.

These results can be interpreted along the lines of a Bayesian framework, closely related to reliability-based cue weighting and predictive coding. Following this rationale, the fronto-parietal network would compute a Bayesian prior and communicate the outcome to early auditory regions. Here, perception is determined on the basis of this prior and incoming sensory evidence (the likelihood function). Since, in the case of recalibration, sensory evidence is ambiguous, perception is mostly determined by the prior. Hence, perception of the A? stimuli follows the audiovisual exposure blocks- /ada/ for A?Vd and /aba/ for A?Vb.

In a separate study, a modeling approach applied to the data from Vroomen et al. (2007) showed that a Bayesian model could explain the behavioral data reflecting both the phenomena of phonetic recalibration and selective speech adaptation (Kleinschmidt and Jaeger 2011). The authors suggest that this "belief-updating model" could provide a unified explanation for both phenomena. The results discussed before, however, demonstrate that the neural underpinnings of both phenomena appear to be distinct, suggesting that separate mechanisms are at play. In future, it will be essential to devise a single coherent model that can explain, both, the phenomena of cross-modal recalibration and selective speech adaptation, while taking their neural mechanisms into account, i.e., an ecologically valid Bayesian model of phonemic aftereffects.

References

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology, 14*, 257–262.

Anstis, S., Verstraten, F. A., & Mather, G. (1998). The motion aftereffect. *Trends in Cognitive Sciences, 2*, 111–117.

Arnal, L. H., & Giraud, A. L. (2012). Cortical oscillations and sensory predictions. *Trends in Cognitive Sciences, 16*, 390–398.

Arnal, L. H., Wyart, V., & Giraud, A.-L. (2011). Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nature Neuroscience, 14*, 797–801.

Beauchamp, M. S. (2005). Statistical criteria in FMRI studies of multisensory integration. *Neuroinformatics, 3*, 93–113.

Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H., & Martin, A. (2004a). Unraveling multisensory integration: Patchy organization within human STS multisensory cortex. *Nature Neuroscience, 7*, 1190–1192.

Beauchamp, M. S., Lee, K. E., Argall, B. D., & Martin, A. (2004b). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron, 41*, 809–823.

Beauchamp, M. S., Nath, A. R., & Pasalar, S. (2010). fMRI-guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect. *Journal of Neuroscience, 30*, 2414–2417.

Bermant, R. I., & Welch, R. B. (1976). Effect of degree of separation of visual-auditory stimulus and eye position upon spatial interaction of vision and audition. *Perceptual and Motor Skills, 43*, 487–493.

Bertelson, P., & Aschersleben, G. (1998). Automatic visual bias of perceived auditory location. *Psychonomic Bulletin & Review, 5*, 482–489.

Bertelson, P., Frissen, I., Vroomen, J., & De Gelder, B. (2006). The aftereffects of ventriloquism: Patterns of spatial generalization. *Attention, Perception, & Psychophysics, 68*, 428–436.

Bertelson, P., & Radeau, M. (1981). Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. *Attention, Perception, & Psychophysics, 29*, 578–584.

Bertelson, P., Vroomen, J. & De Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk aftereffect. *Psychological Sciences*, 14, 592–597.

Besle, J., Bertrand, O., & Giard, M.-H. (2009). Electrophysiological (EEG, sEEG, MEG) evidence for multiple audiovisual interactions in the human auditory cortex. *Hearing Research, 258*, 143–151.

Besle, J., Fischer, C., Bidet-Caulet, A., Lecaignard, F., Bertrand, O., & Giard, M.-H. (2008). Visual activation and audiovisual interactions in the auditory cortex during speech perception: Intracranial recordings in humans. *Journal of Neuroscience, 28*, 14301–14310.

Besle, J., Fort, A., Delpuech, C., & Giard, M. H. (2004). Bimodal speech: Early suppressive visual effects in human auditory cortex. *European Journal of Neuroscience, 20*, 2225–2234.

Blank, H., Anwander, A., & von Kriegstein, K. (2011). Direct structural connections between voice-and face-recognition areas. *Journal of Neuroscience, 31*, 12906–12915.

Callan, D. E., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2001). Multimodal contribution to speech perception revealed by independent component analysis: A single-sweep EEG case study. *Cognitive Brain Research, 10*, 349–353.

Callan, D. E., Jones, J. A., Munhall, K., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *Neuroreport, 14*, 2213–2218.

Callan, D. E., Jones, J. A., Munhall, K., Kroos, C., Callan, A. M., & Vatikiotis-Bateson, E. (2004). Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information. *Journal of Cognitive Neuroscience, 16*, 805–816.

Calvert, G., Spence, C., & Stein, B. E. (Eds.) (2004). *The handbook of multisensory processes*. Cambridge, MA: MIT Press.

Calvert, G. A. (2001). Crossmodal processing in the human brain: Insights from functional neuroimaging studies. *Cerebral Cortex, 11*, 1110–1123.

Calvert, G. A., Brammer, M. J., Bullmore, E. T., Campbell, R., Iversen, S. D., & David, A. S. (1999). Response amplification in sensory-specific cortices during crossmodal binding. *Neuroreport, 10*, 2619–2623.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology, 10*, 649–657.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., et al. (1997). Activation of auditory cortex during silent lipreading. *Science, 276*, 593–596.

Campbell, R., & Capek, C. (2008). Seeing speech and seeing sign: Insights from a fMRI study. *International Journal of Audiology, 47*, S3–S9.

Campbell, R., MacSweeney, M., Surguladze, S., Calvert, G., McGuire, P., Suckling, J., et al. (2001). Cortical substrates for the perception of face actions: An fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning). *Cognitive Brain Research, 12*, 233–243.

Capek, C. M., Bavelier, D., Corina, D., Newman, A. J., Jezzard, P., & Neville, H. J. (2004). The cortical organization of audio-visual sentence comprehension: An fMRI study at 4 Tesla. *Cognitive Brain Research, 20*, 111–119.

Cappe, C., Rouiller, E. M., & Barone, P. (2009). Multisensory anatomical pathways. *Hearing Research, 258*, 28–36.

Cappe, C., Thut, G., Romei, V., & Murray, M. M. (2010). Auditory–visual multisensory interactions in humans: Timing, topography, directionality, and sources. *Journal of Neuroscience, 30*, 12572–12580.

Clavagnier, S., Falchier, A., & Kennedy, H. (2004). Long-distance feedback projections to area V1: Implications for multisensory integration, spatial awareness, and visual consciousness. *Cognitive, Affective, & Behavioral Neuroscience, 4*, 117–126.

Clos, M., Langner, R., Meyer, M., Oechslin, M. S., Zilles, K., & Eickhoff, S. B. (2012). Effects of prior information on decoding degraded speech: An fMRI study. *Human Brain Mapping, 35*, 61–74.

Diehl, R. L. (1981). Feature detectors for speech: A critical reappraisal. *Psychological Bulletin, 89*, 1.

Diehl, R. L., Elman, J. L., & McCusker, S. B. (1978). Contrast effects on stop consonant identification. *Journal of Experimental Psychology: Human Perception and Performance, 4*, 599.

Driver, J., & Noesselt, T. (2008). Multisensory interplay reveals crossmodal influences on 'sensory-specific' brain regions, neural responses, and judgments. *Neuron, 57*, 11–23.

Eimas, P. D., & Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology, 4*, 99–109.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*, 429–433.

Falchier, A., Clavagnier, S., Barone, P., & Kennedy, H. (2002). Anatomical evidence of multimodal integration in primate striate cortex. *Journal of Neuroscience, 22*, 5749–5759.

Falchier, A., Schroeder, C. E., Hackett, T. A., Lakatos, P., Nascimento-Silva, S., Ulbert, I., et al. (2010). Projection from visual areas V2 and prostriata to caudal auditory cortex in the monkey. *Cerebral Cortex, 20*, 1529–1538.

Fetsch, C. R., Pouget, A., DeAngelis, G. C., & Angelaki, D. E. (2012). Neural correlates of reliability-based cue weighting during multisensory integration. *Nature Neuroscience, 15*, 146–154.

Formisano, E., De Martino, F., Bonte, M., & Goebel, R. (2008). "Who" is saying "what"? Brain-based decoding of human voice and speech. *Science, 322*, 970–973.

Forster, B., Cavina-Pratesi, C., Aglioti, S. M., & Berlucchi, G. (2002). Redundant target effect and intersensory facilitation from visual-tactile interactions in simple reaction time. *Experimental Brain Research, 143*, 480–487.

Frassinetti, F., Bolognini, N., & Làdavas, E. (2002). Enhancement of visual perception by cross-modal visuo-auditory interaction. *Experimental Brain Research, 147*, 332–343.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 360*, 815–836.

Ghazanfar, A. A. (2012). Unity of the senses for primate vocal communication. In M. T. Murray & M. M. Wallace (Eds.), *The neural bases of multisensory processes* (pp. 653–666). Boca Raton: CRC.

Ghazanfar, A. A., Chandrasekaran, C., & Logothetis, N. K. (2008). Interactions between the superior temporal sulcus and auditory cortex mediate dynamic face/voice integration in rhesus monkeys. *Journal of Neuroscience, 28*, 4457–4469.

Ghazanfar, A. A., Maier, J. X., Hoffman, K. L., & Logothetis, N. K. (2005). Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *Journal of Neuroscience, 25*, 5004–5012.

Ghazanfar, A. A., & Schroeder, C. E. (2006). Is neocortex essentially multisensory? *Trends in Cognitive Sciences, 10*, 278–285.

Giard, M. H., & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study. *Journal of Cognitive Neuroscience, 11*, 473–490.

Gibson, J. J., & Radner, M. (1937). Adaptation, after-effect and contrast in the perception of tilted lines. I. Quantitative studies. *Journal of Experimental Psychology, 20*, 453.

Gilbert, C. D., Sigman, M., & Crist, R. E. (2001). The neural basis of perceptual learning. *Neuron, 31*, 681–697.

Goebel, R., Esposito, F., & Formisano, E. (2006). Analysis of functional image analysis contest (FIAC) data with brainvoyager QX: From single-subject to cortically aligned group general linear model analysis and self-organizing group independent component analysis. *Human Brain Mapping, 27*, 392–401.

Gondan, M., & Röder, B. (2006). A new method for detecting interactions between the senses in event-related potentials. *Brain Research, 1073*, 389–397.

Hackett, T. A., De La Mothe, L. A., Ulbert, I., Karmos, G., Smiley, J., & Schroeder, C. E. (2007). Multisensory convergence in auditory cortex. II. Thalamocortical connections of the caudal superior temporal plane. *Journal of Comparative Neurology, 502*, 924–952.

Hall, D. A., Fussell, C., & Summerfield, A. Q. (2005). Reading fluent speech from talking faces: Typical brain networks and individual differences. *Journal of Cognitive Neuroscience, 17*, 939–953.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293*, 2425–2430.

Haynes, J.-D., & Rees, G. (2005a). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience, 8*, 686–691.

Haynes, J.-D., & Rees, G. (2005b). Predicting the stream of consciousness from activity in human visual cortex. *Current Biology, 15*, 1301–1307.

Hughes, H. C., Reuter-Lorenz, P. A., Nozawa, G., & Fendrich, R. (1994). Visual-auditory interactions in sensorimotor processing: Saccades versus manual responses. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 131–153.

Jiang, W., Wallace, M. T., Jiang, H., Vaughan, J. W., & Stein, B. E. (2001). Two cortical areas mediate multisensory integration in superior colliculus neurons. *Journal of Neurophysiology, 85*, 506–522.

Kayser, C., Petkov, C. I., Augath, M., & Logothetis, N. K. (2005). Integration of touch and sound in auditory cortex. *Neuron, 48*, 373–384.

Kayser, J., Tenke, C. E., Gates, N. A., & Bruder, G. E. (2007). Reference-independent ERP old/new effects of auditory and visual word recognition memory: Joint extraction of stimulus-and response-locked neuronal generator patterns. *Psychophysiology, 44*, 949–967.

Kilian-Hütten, N., Valente, G., Vroomen, J., & Formisano, E. (2011a). Auditory cortex encodes the perceptual interpretation of ambiguous sound. *Journal of Neuroscience, 31*, 1715–1720.

Kilian-Hütten, N., Vroomen, J., & Formisano, E. (2011b). Brain activation during audiovisual exposure anticipates future perception of ambiguous speech. *Neuroimage, 57*, 1601–1607.

Kleinschmidt, D., & Jaeger, T. F. (2011). A Bayesian belief updating model of phonetic recalibration and selective adaptation. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics* (pp. 10–19). Association for Computational Linguistics.

Klemm, O. (1909). Lokalisation von Sinneseindrücken bei disparaten Nebenreizen. [Localization of sensory impressions with disparate distractors]. *Psychologische Studien (Wundt) 5*, 73–161.

Kuhl, P. K., & Meltzoff, A. N. (1982). *The bimodal perception of speech in infancy*. Washington, DC: American Association for the Advancement of Science.

Lakatos, P., Chen, C.-M., O'Connell, M. N., Mills, A., & Schroeder, C. E. (2007). Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron, 53*, 279–292.

Laurienti, P. J., Perrault, T. J., Stanford, T. R., Wallace, M. T., & Stein, B. E. (2005). On the use of superadditivity as a metric for characterizing multisensory integration in functional neuroimaging studies. *Experimental Brain Research, 166*, 289–297.

Ludman, C., Summerfield, A. Q., Hall, D., Elliott, M., Foster, J., Hykin, J. L., et al. (2000). Lip-reading ability and patterns of cortical activation studied using fMRI. *British Journal of Audiology, 34*, 225–230.

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., & Parra, L. C. (2009). Lip-reading aids word recognition most in moderate noise: A Bayesian explanation using high-dimensional feature space. *PLoS One, 4*, e4638.

Macleod, A., & Summerfield, Q. (1990). A procedure for measuring auditory and audiovisual speech-reception thresholds for sentences in noise: Rationale, evaluation, and recommendations for use. *British Journal of Audiology, 24*, 29–43.

MacSweeney, M., Woll, B., Campbell, R., McGuire, P. K., David, A. S., Williams, S. C., et al. (2002). Neural systems underlying British Sign Language and audio-visual English processing in native users. *Brain, 125*, 1583–1593.

McGettigan, C., Faulkner, A., Altarelli, I., Obleser, J., Baverstock, H., & Scott, S. K. (2012). Speech comprehension aided by multiple modalities: Behavioural and neural interactions. *Neuropsychologia, 50*, 762–776.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746–748

Middelweerd, M., & Plomp, R. (1987). The effect of speechreading on the speech-reception threshold of sentences in noise. *The Journal of the Acoustical Society of America, 82*, 2145–2147.

Miller, L. M., & D'esposito, M. (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *Journal of Neuroscience, 25*, 5884–5893.

Molholm, S., Ritter, W., Murray, M. M., Javitt, D. C., Schroeder, C. E., & Foxe, J. J. (2002). Multisensory auditory–visual interactions during early sensory processing in humans: A high-density electrical mapping study. *Cognitive Brain Research, 14*, 115–128.

Möttönen, R., Schürmann, M., & Sams, M. (2004). Time course of multisensory interactions during audiovisual speech perception in humans: A magnetoencephalographic study. *Neuroscience Letters, 363*, 112–115.

Munhall, K. G., & Buchan, J. N. (2004). Something in the way she moves. *Trends in Cognitive Science, 8*, 51–53.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility head movement improves auditory speech perception. *Psychological Science, 15*, 133–137.

Murray, M. T., & Wallace, M. M. (Eds.) (2011). *The neural bases of multisensory processes*. Boca Raton: CRC.

Myers, E. B., Blumstein, S. E., Walsh, E., & Eliassen, J. (2009). Inferior frontal regions underlie the perception of phonetic category invariance. *Psychological Science, 20*, 895–903.

Nath, A. R., & Beauchamp, M. S. (2011). Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. *Journal of Neuroscience, 31*, 1704–1714.

Nath, A. R., Fava, E. E., & Beauchamp, M. S. (2011). Neural correlates of interindividual differences in children's audiovisual speech perception. *Journal of Neuroscience, 31*, 13963–13971.

Naumer, M. J., Doehrmann, O., Müller, N. G., Muckli, L., Kaiser, J., & Hein, G. (2009). Cortical plasticity of audio–visual object representations. *Cerebral Cortex, 19*, 1641–1653.

Noesselt, T., Rieger, J. W., Schoenfeld, M. A., Kanowski, M., Hinrichs, H., Heinze, H.-J., et al. (2007). Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices. *Journal of Neuroscience, 27*, 11431–11441.

Noppeney, U., Josephs, O., Hocking, J., Price, C. J., & Friston, K. J. (2008). The effect of prior visual information on recognition of speech and sounds. *Cerebral Cortex, 18*, 598–609.

Ojanen, V., Möttönen, R., Pekkola, J., Jääskeläinen, I. P., Joensuu, R., Autti, T., et al. (2005). Processing of audiovisual speech in Broca's area. *Neuroimage, 25*, 333–338.

Olson, I. R., Gatenby, J. C., & Gore, J. C. (2002). A comparison of bound and unbound audio–visual information processing in the human cerebral cortex. *Cognitive Brain Research, 14*, 129–138.

Ozker, M., Schepers, I. M., Magnotti, J. F., Yosher, D., & Beauchamps, M. (2017). A double dissociation between anterior and posterior superior temporal gyrus for processing audiovi-

sual speech demonstrated by electrocorticography. *Journal of Cognitive Neuroscience, 2916*, 1044–1060.

Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I. P., Möttönen, R., Tarkiainen, A., et al. (2005). Primary auditory cortex activation by visual speech: An fMRI study at 3 T. *Neuroreport, 16*, 125–128.

Peelle, J.E. & Sommers, M.S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex, 68*, 169–181.

Ponton, C. W., Bernstein, L. E., & Auer, E. T. (2009). Mismatch negativity with visual-only and audiovisual speech. *Brain Topography, 21*, 207–215.

Purkinje, J. (1820). Beiträge zur näheren Kenntnis des Schwindels. *Med Jahrb kuk Staates (Wien), 6*, 23–35.

Radeau, M., & Bertelson, P. (1974). The after-effects of ventriloquism. *The Quarterly Journal of Experimental Psychology, 26*, 63–71.

Radeau, M., & Bertelson, P. (1977). Adaptation to auditory-visual discordance and ventriloquism in semirealistic situations. *Perception & Psychophysics, 22*, 137–146.

Raizada, R. D., & Poldrack, R. A. (2007). Selective amplification of stimulus differences during categorical processing of speech. *Neuron, 56*, 726–740.

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 97–113). London: Erlbaum.

Remez, R. E. (2012). Three puzzles of multimodal speech perception. In G. Bailly, P. Perrier, & E. Vatikiotis-Bateson (Eds.), *Audiovisual speech processing* (pp. 4–20). Cambridge: Cambridge University Press.

Roberts, M., & Summerfield, Q. (1981). Audiovisual presentation demonstrates that selective adaptation in speech perception is purely auditory. *Attention, Perception, & Psychophysics, 30*, 309–314.

Rosenblum, L. D., Pisoni, D., & Remez, R. (2005). Primacy of multimodal speech perception. In D. Pisoni & R. Remez (Eds.) *Handbook of speech perception* (pp. 51–78). Malden: Blackwell.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex, 17*, 1147–1153.

Samuel, A. G. (1986). Red herring detectors and speech perception: In defense of selective adaptation. *Cognitive Psychology, 18*, 452–499.

Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., & Puce, A. (2008). Neuronal oscillations and visual amplification of speech. *Trends in Cognitive Sciences, 12*, 106–113.

Schroeder, C. E., Smiley, J., Fu, K. G., McGinnis, T., O'Connell, M. N., & Hackett, T. A. (2003). Anatomical mechanisms and functional implications of multisensory convergence in early cortical processing. *International Journal of Psychophysiology, 50*, 5–17.

Schröger, E., & Widmann, A. (1998). Speeded responses to audiovisual signal changes result from bimodal integration. *Psychophysiology, 35*, 755–759.

Schwiedrzik, C. M., Ruff, C. C., Lazar, A., Leitner, F. C., Singer, W., & Melloni, L. (2014). Untangling perceptual memory: Hysteresis and adaptation map into separate cortical networks. *Cerebral Cortex, 24*, 1152–1164.

Sekuler, R., Sekuler, A. B., & Lau, R. (1997). Sound alters visual motion perception. *Nature, 385*, 308.

Senkowski, D., Talsma, D., Grigutsch, M., Herrmann, C. S., & Woldorff, M. G. (2007). Good times for multisensory integration: Effects of the precision of temporal synchrony as revealed by gamma-band oscillations. *Neuropsychologia, 45*, 561–571.

Sheppard, J. P., Raposo, D., & Churchland, A. K. (2013). Dynamic weighting of multisensory stimuli shapes decision-making in rats and humans. *Journal of Vision, 13*, 4.

Shimojo, S., & Shams, L. (2001). Sensory modalities are not separate modalities: Plasticity and interactions. *Current Opinion in Neurobiology, 11*, 505–509.

Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2005). Listening to talking faces: Motor cortical activation during speech perception. *Neuroimage, 25*, 76–89.

Sohoglu, E., Peelle, J. E., Carlyon, R. P., & Davis, M. H. (2012). Predictive top-down integration of prior knowledge during speech perception. *Journal of Neuroscience, 32*, 8443–8453.

Staeren, N., Renvall, H., De Martino, F., Goebel, R., & Formisano, E. (2009). Sound categories are represented as distributed patterns in the human auditory cortex. *Current Biology, 19*, 498–502.

Stein, B., & Meredith, M. (1990). Multimodal integration: Neural and behavioral solutions for dealing with stimuli from different modalities. *Annals of the New York Academy of Science, 606*, 51–70.

Stein, B. E., Huneycutt, W. S., & Meredith, M. A. (1988). Neurons and behavior: The same rules of multisensory integration apply. *Brain Research, 448*, 355–358.

Stein, B. E., Stanford, T. R., Ramachandran, R., Perrault, T. J., & Rowland, B. A. (2009). Challenges in quantifying multisensory integration: Alternative criteria, models, and inverse effectiveness. *Experimental Brain Research, 198*, 113.

Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience, 19*, 1964–1973.

Stevenson, R. A., Geoghegan, M. L., & James, T. W. (2007). Superadditive BOLD activation in superior temporal sulcus with threshold non-speech objects. *Experimental Brain Research, 179*, 85–95.

Stevenson, R. A., & James, T. W. (2009). Audiovisual integration in human superior temporal sulcus: Inverse effectiveness and the neural processing of speech and object recognition. *Neuroimage, 44*, 1210–1223.

Stratton, G. M. (1897). Vision without inversion of the retinal image. *Psychological Review, 4*, 341.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America, 26*, 212–215.

Teder-Sälejärvi, W., McDonald, J., Di Russo, F., & Hillyard, S. (2002). An analysis of audio-visual crossmodal integration by means of event-related potential (ERP) recordings. *Cognitive Brain Research, 14*, 106–114.

Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America, 102*, 1181–1186.

von Kriegstein, K. (2012). A multisensory perspective on human auditory communication. In M. T. Murray & M. M. Wallace (Eds.), *The neural bases of multisensory processes* (pp. 683–702). Boca Raton: CRC.

Vroomen, J., & Baart, M. (2009). Phonetic recalibration only occurs in speech mode. *Cognition, 110*, 254–259.

Vroomen, J., & Baart, M. (2012). Phonetic recalibration in audiovisual speech. In M. T. Murray & M. M. Wallace (Eds.), *The neural bases of multisensory processes* (pp. 363–380). Boca Raton: CRC.

Vroomen, J., & de Gelder, B. (2004). Temporal ventriloquism: Sound modulates the flash-lag effect. *Journal of Experimental Psychology: Human Perception and Performance, 30*, 513–518.

Vroomen, J., & Stekelenburg, J. J. (2010). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. *Journal of Cognitive Neuroscience, 22*, 1583–1596.

Vroomen, J., van Linden, S., De Gelder, B., & Bertelson, P. (2007). Visual recalibration and selective adaptation in auditory–visual speech perception: Contrasting build-up courses. *Neuropsychologia, 45*, 572–577.

Vroomen, J., van Linden, S., Keetels, M., De Gelder, B., & Bertelson, P. (2004). Selective adaptation and recalibration of auditory speech by lipread information: Dissipation. *Speech Communication, 44*, 55–61.

Watkins, K. E., Strafella, A. P., & Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia, 41*, 989–994.

Wright, T. M., Pelphrey, K. A., Allison, T., McKeown, M. J., & McCarthy, G. (2003). Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cerebral Cortex, 13*, 1034–1043.

Yehia, H. C., Kuratate, T., & Vatikiotis-Bateson, E. (2002). Linking facial animation, head motion and speech acoustics. *Journal of Phonetics, 30*, 555–568.